Informatie zoeken, vinden en vindbaar maken in vijf thema's

Van **blob** tot **bot**

In de relatief coronamaatregelen-vrije *window* in oktober was er na anderhalf jaar weer een live VOGIN-IP-lezing in de Openbare Bibliotheek Amsterdam (OBA). De achtste editie van het gezamenlijke evenement van Stichting VOGIN en vakblad IP over zoeken, vinden en vindbaar maken van informatie. Een terugblik in vijf thema's.

vogin IP vakblad voor informatieprofessionals

Marjo Bakker
Teamleider Collecties, data-steward en vakreferent bij het NIOD Instituut voor Oorlogs-, Holocaust- en Genocidestudies

Cindy Lammers
Hoofdredacteur van IP

Koffie, koek en knikkende hoofden. Ja, leuk om elkaar weer te zien en te spreken, hoor je lachend om je heen op de bovenste etage van de OBA in Amsterdam. Op 21 oktober kon dat opeens weer. Met coronatoegangsbewijs, zonder mondkapje en anderhalve meter. Vorig jaar viel de fysieke VOGIN-IP-lezing in het water door de eerste lockdown op 15 maart, vier dagen voor het event zou plaatsvinden. De organisatie tuigde daarna wel een onlineprogramma op, dat goed werd 'bezocht', maar anderhalf jaar verder is het live weer mogelijk. Wel blijkt vlak voor aanvang dat dagvoorzitter Annemarie van Campen door quarantaine moet afhaken. Gelukkig is oude bekende Peter van Gorsel bereid haar taak over te nemen. Hij leidt, af en toe improviserend, de aanwezigen door een dag met acht lezingen en evenzoveel workshops.

Wat opvalt aan het eind van deze dag propvol informatie is een zekere bescheidenheid bij de twee keynotesprekers: 'We zijn er nog lang niet', en: 'Het is nog best moeilijk.' Daarnaast is er enthousiasme bij alle sprekers om tips en tricks te delen.

— THEMA:

FAKE

De eerste keynote, door Zeno Geradts (forensisch onderzoeker bij het Nederlands Forensisch Instituut en bijzonder hoogleraar Forensic Data Science aan de UvA), leert ons dat zelf deepfakes maken redelijk makkelijk is. Bijvoorbeeld via DeepFaceLab. Lastiger is het om te beoordelen of een foto of video nep is of niet. Elke keer nadat een artikel is gepubliceerd over hoe je een deepfake kunt herkennen, worden de deepfakes weer lastiger te detecteren omdat de makers van deepfakesoftware deze detectiemethoden ook goed bestuderen. Het is een 'arms race', aldus Geradts.
Bij forensisch onderzoek komt ook vaak de andere kant aan de orde. Wat

Foto's: Eef Evers

Zeno Geradts

de MH17-ramp betreft beweerde een Russische expert bijvoorbeeld dat de beelden gemanipuleerd waren. Dat dit niet zo is, valt moeilijk te bewijzen. Het is vaak makkelijker aan te tonen dat iets nep is dan dat iets echt is, weet Geradts. Ook de zaal heeft het lastig als hij van enkele getoonde filmpjes vraagt of het gaat om deepfake of echt. Gelukkig zijn er steeds meer bedrijven die zich bezighouden met deepfakedetectie, ook in Nederland, waaronder DuckDuckGoose en Sensity. Verder is er samenwerking met de UvA om AI-technieken bij het herkennen van deepfakes te testen. Desalniettemin begrijpt hij de angst, bijvoorbeeld als een nep-Mark Rutte iets zou zeggen wat mensen zouden gaan opvolgen. 'Het blijft een kat-en-muisspel. Ik ben eerlijk over hoe moeilijk het is.'

— THEMA:

DE GEVAREN VAN CYBERSPACE

Militair, academicus en softwarespecialist Jelle van Haaster schetst in korte tijd een beeld van de exponentiële acceptatie en adaptatie van technologie, en hoe internet onderdelen van de samenleving heeft veranderd. Dus ook oorlogsvoering – als onderdeel van die samenleving. 'Er is sprake van meer connectiviteit dan ooit. Dit leidt tot fundamentele verschuivingen van (staats)macht. Neem Puerto Rico, dat in 2017 hulp nodig had na de orkaanramp en Tesla hiervoor benaderde. Deze hulp kwam, niet vanuit de goedheid van Elon Musks hart, maar omdat dit goed is voor zijn reputatie en positie.'

Jelle van Haaster

'Who codes matters' is nog zoiets, vervolgt Van Haaster. 'Technologie is niet neutraal. Er zitten stereotypen in software ingebakken, simpelweg omdat programmeurs ook maar mensen zijn. En denk aan Hypponen's law: if it's smart, it's vulnerable.' Wat Defensie betreft is er een geheel ander speelveld ontstaan, weet hij: het is nu land, zee, lucht én (cyber)space. Waarom zou je nog fysiek naar een conflictgebied gaan als je iedereen over de hele wereld kunt bereiken? Het is ook veel onduidelijker geworden waar wel of niet een conflict is, zie de Krim en de aanvallen van hackers. Vechten in de eenentwintigste eeuw is tricky geworden.' Gelukkig, besluit Van Haaster, staat Nederland in de cyberpowerranglijst op nummer 5, na de VS, China, het VK en Rusland.

'Waarom zou je nog fysiek naar een conflictgebied gaan als je iedereen over de hele wereld kunt bereiken?'

'Ik zeg: geen datalake of AI, maar gewoon handwerk en structuur om een datakerkhof te voorkomen'

— THEMA:

BLOCK THE BLOB, OFTEWEL: HET BELANG VAN GESTRUCTUREERDE INFORMATIE

Contentstrateeg Ilse Jonker introduceert de term 'blob', ongestructureerde data in één veld van een datamodel gepropt, of content zonder metadata, en die term resoneert in de andere lezingen. Blob belemmert de vindbaarheid van informatie, terwijl we die juist zo hoog in het vaandel hebben staan. Hoe komt dat? Jonker geeft enkele mooie voorbeelden. Het heeft te maken met dat je te graag te snel informatie online zet, 'informatieliefde' in haar jargon, en dat je verouderde informatie of websites niet op-

Ilse Jonker

ruimt. 'Doe bijvoorbeeld eens de zoekactie site:[websitenaam.nl]', suggereert ze. 'Dan zie je hoeveel pagina's jouw domein heeft. Klopt dat ongeveer met wat je denkt dat je hebt? Deze verwaarlozing, in combinatie met informatieliefde en blob, is het recept voor rampspoed.'
Verder heb je binnen je organisatie misschien te maken met mensen die zeggen: maar we hebben toch een chatbox, datalake, knowledge graph of AI? 'Ja, klopt, maar dat ontslaat je niet van het structureren van je content', weet Jonker. 'Dat is vaak lastiger dan gedacht; data zijn niet zo gestructureerd. Dus ik zeg: geen datalake of AI, maar gewoon handwerk en structuur om een datakerkhof te voorkomen.'
Gelukkig zijn er tips. Ten eerste: niet alles hoeft digitaal, en ten tweede:

houd de eindgebruiker in gedachten. Verder: kruip uit je silo, zoek coalities. De vierde tip luidt: maak het chefsache om het beter te organiseren, bijvoorbeeld via visualisatie van de kosten. En als laatste tip: houd rekening met een lange adem, dan valt het altijd mee. 'De verandering van zoiets ogenschijnlijk simpels als het invoeren van nieuwe documentnamen kostte een verzekeringsmaatschappij niet minder dan twee jaar.'

Linked open data

Gestructureerde data zijn ook het thema van de lezing van Lizzy Jongma (ICT-projectleider bij het Netwerk Oorlogsbronnen (NOB): 'LOD voor WO2'. Het NOB verbindt bronnen over de Tweede Wereldoorlog digitaal met elkaar, zodat de bronnen

Lizzy Jongma

vindbaar en bruikbaar worden en in samenhang en met context kunnen worden gepresenteerd. De technieken om deze bronnen op te halen, werken alleen niet altijd even goed, vindt Jongma. Ze fulmineert op grappige wijze tegen de onvolkomenheden van de diverse technieken, zoals OAI-PMH, Dublin Core en SPARQL endpoint, maar is blij dat de data inmiddels worden gedeeld, ook al is die informatie dan soms als blob verpakt. Het mooie is dat door het bij elkaar halen van veel data, nieuwe kennis en inzichten ontstaan, vertelt ze. Bijvoorbeeld dat Rotterdam in de oorlog liquidatiehoofdstad van Nederland was. Of dat de piek in het sterftecijfer van de oorlog niet aan het begin en eind van de oorlog lag, maar in het midden toen de meeste Joden in concentratiekampen zijn vermoord. 'Snippers van mensenlevens verspreid over verschillende archieven komen nu bij elkaar met dank aan een geautomatiseerd proces. Wie bij wie in de deportatietrein zat, staat niet in een archiefstuk – al is de Tweede Wereldoorlog waanzinnig goed gedocumenteerd. Je moet data combineren, en het NOB doet dat via linked open data (LOD) en zoekstrategieën. Zo kom je van individuele mensen bij big data uit.'

Semantisaurus

Het NOB hanteert LOD inmiddels op zijn eigen manier. Het streven is LOD bij de bron, maar zolang dat er bij erfgoedinstellingen nog niet is, maakt het NOB zelf de LOD en levert deze weer terug. Contextualiseren van de data gaat via het oeroude middel van de thesaurus, bij het NOB inmiddels uitgebouwd tot een mini-encyclopedie. Jongma spreekt graag over de 'semantisaurus', waarin allerlei relaties zijn gelegd tussen mensen, plekken en gebeurtenissen.

Constant Hijzen

Kennis van zoektechnologie is nog heel beperkt bij erfgoedprofessionals, weet ze. 'We gooien het ongestructureerd op internet en hopen dat mensen dan weten hoe ze moeten zoeken en vinden. Ik breek daarom een lans voor meer kennis over wat je nog meer kunt doen dan alleen dat zoekvakje op de website. Van blobs naar ballen – de zoekstrategieën bij het NOB worden gevisualiseerd als ballen, vandaar. Waarom zou je moeten weten dat je met een asterisk moet zoeken, waarom niet gewoon in natuurlijke taal?'

'Kennis van zoektechnologie is nog heel beperkt bij erfgoedprofessionals, veel gooien we nog ongestructureerd op internet'

Jerry Vermanen en Peter van Gorsel

'Een tip om verstopte data te vinden is: plan een koffieafspraak, want informatie komt bij mensen vandaan'

THEMA:

HOE VIND JE INFORMATIE DIE NIET GEVONDEN WIL WORDEN?

Constant Hijzen (research fellow aan het Institute of Security and Global Affairs van de Universiteit Leiden) laat zien hoe je geheime informatie toch vindt. Voor zijn onderzoek naar de inlichtingen- en veiligheidsdiensten heeft hij geen geheim archief kunnen raadplegen, maar hij wist toch aan relevante informatie te komen. Hoe je hiervoor te werk kunt gaan – gesneden koek voor historici en archivarissen – heeft te maken met context en bronnenkritiek, weet hij. 'Bedenk wie de spelers zijn op je onderwerpsgebied en wie de ontvangers zijn van, in mijn geval, inlichtingenproducten, en ga in de archieven van die mensen of organisaties speuren naar relevante brieven en mailwisselingen en nota's. Dan kun je alsnog veel te weten komen.'

Leer scrapen

Ook Jerry Vermanen (datajournalist bij KRO-NCRV's *Pointer* en coauteur van *Handboek Internetresearch en datajournalistiek*) heeft tips om verstopte data te vinden. Plan een koffieafspraak, stelt hij, want informatie komt bij mensen vandaan. Maak je eigen data-alarmeringsdienst: 'Gooi bijvoorbeeld zoekopdrachten naar filetypes, in combinatie trefwoorden en sites waar je informatie verwacht, in Google Alerts.' Nog een tip is: stel zelf datasets samen – 'Dat mag gewoon in een spreadsheet' – en leer scrapen. Ook je telefoon kan interessante info bieden, weet Vermanen. 'Het is je beste logboek voor locatiegegevens, maar ook voor bijvoorbeeld het exporteren van WhatsApp- of Telegram-gesprekken.'

De lezing van Edgar Meij (hoofd Artificial Intelligence (AI) Discovery group bij Bloomberg Engineering) gaat over zoeken en vinden van informatie in de financiële wereld. Bloomberg zet hierbij AI in via het ondersteunen van natuurlijke taal in de Bloomsbury terminal. Je stelt een vraag in natuurlijke taal, bijvoorbeeld: What is the yearly net income of Google and IBM in the last 20 years?', en krijgt een antwoord dat automatisch wordt gegenereerd, maar Bloomberg kijkt ook naar hoe via serendipiteit een antwoord op vragen kan worden geven. Het bedrijf hanteert hiervoor een eigen taxonomie.

Edgar Meij

Verder analyseert het onderwerpen die op Twitter en nieuwssites langskomen. Over Amazon bijvoorbeeld komen dagelijks circa 174 artikelen voorbij, vertelt Meij. 'Hoe analyseer je die? Hoe krijg je dit als pakketje bij de mensen die bijvoorbeeld bij een pensioenfonds werken? Hoe cluster je, hoe genereer je een samenvatting – in real time? Verder kan nieuws voorspellen wat er met aandelenkoersen gebeurt. Het is een uitdaging om die informatie goed bij de klanten te brengen en machine readable te maken.' Meer over AI bij Bloomberg vind je op techatbloomberg.com/AI/.

Taalmodel GPT-3

De dag wordt afgesloten met de keynote van Antal van den Bosch (directeur van het KNAW Meertens Instituut en bijzonder hoogleraar Taal en Kunstmatige Intelligentie aan de UvA), die ingaat op de vraag of neurale taalmodellen de toekomst van AI zijn of slechts een goede goocheltruc. Een echte AI'er herken je doordat hij zegt: 'We zijn er nog lang niet.' Dat staat misschien in contrast met de hype die om AI heen hangt en met de sterke staaltjes AI die Van den Bosch laat zien. Bijvoorbeeld GPT-3 van het bedrijf OpenAI (opgericht door Elon Musk, maar die trok zich later terug). GPT-3 is een taalmodel dat in staat is om teksten af te maken, vragen te beantwoorden en professionele teksten om te zetten in teksten voor leken. 'Hiervoor is een "mentaal woordenboek" nodig; hoe woorden bij elkaar vóórkomen', zegt Van den Bosch. 'We zeggen bijvoorbeeld "sterke koffie" en

Antal van den Bosch

niet "krachtige koffie". Als je een taalmodel wilt maken, moet de kennis uit het mentale woordenboek erin zitten. Google Translate bijvoorbeeld heeft dat mentale woordenboek geïncorporeerd. Vanaf 2017 werkt het met *neural machine translation*, dat wil zeggen dat het systeem, getraind op een corpus van vertaalde teksten, weet in welke taal welke combinaties van woorden bij elkaar voorkomen. Google Translate is sindsdien veel beter geworden, ook doordat langeafstandsafhankelijkheden worden overbrugd door meer geheugen.'

'Is het punt bereikt waarop de AI de collectieve intelligentie van de mens passeert? Het antwoord is nee'

Nog geen denkende machine

Een aansprekend voorbeeld van AI dat Van den Bosch laat zien, is AsiBot, de robot die een laatste (nieuw) hoofdstuk aan Isaac Asimovs sciencefictionboek *I, Robot* toevoegde in samenwerking met schrijver Ronald Giphart. AsiBot schreef uiteindelijk 50 procent van de tekst. De robot was getraind op tienduizend Nederlandse romans en kon ook in de stijl van andere auteurs tekst genereren. Hier kun je mooi de show mee stelen, weet Van den Bosch. 'Het gaat richting de denkende machine, maar is dit nu de algemene kunstmatige intelligentie? Is de singulariteit – het punt waarop de AI de collectieve intelligentie van de mens passeert – bereikt? Het antwoord is nee, want het geloofwaardig kunnen vervolgen van een tekst of dialoog, wat AI nu kan, is niet alles wat wij als mensen doen. Onze intelligentie is meer dan goed kunnen kletsen. De menselijke intelligentie bestaat ook uit redeneren, plannen en abstraheren. Daarom is de huidige AI knap, maar nog geen denkende machine.' <

Zeven van de acht lezingen zijn online te bekijken op vogin-ip-lezing.net/2021/11/09/vogin-ip-lezingen-terugkijken/. In de volgende IP meer aandacht voor de workshops tijdens de VOGIN-IP-lezing 2021.